Identifiability and inference of non-parametric rates-across-sites models on large-scale phylogenies *

Elchanan Mossel† Sebastien Roch‡

August 2, 2011

Abstract

Mutation rate variation across loci is well known to cause difficulties, notably identifiability issues, in the reconstruction of evolutionary trees from molecular sequences. Here we introduce a new approach for estimating general rates-across-sites models. Our results imply, in particular, that large phylogenies are typically identifiable under rate variation. We also derive sequence-length requirements for high-probability reconstruction.

Our main contribution is a novel algorithm that clusters sites according to their mutation rate. Following this site clustering step, standard reconstruction techniques can be used to recover the phylogeny. Our results rely on a basic insight: that, for large trees, certain site statistics experience concentration-of-measure phenomena.

*Keywords: phylogenetic reconstruction, rates-across-sites models, concentration of measure.

†U.C. Berkeley and Weizmann Institute of Science. Supported by DMS 0548249 (CAREER) award, by DOD ONR grant N000141110140, by ISF grant 1300/08 and by ERC PIRG04-GA-2008-239137 grant.

‡Department of Mathematics and Bioinformatics Program, UCLA. Work supported by NSF grant DMS-1007144.

1 Introduction

The evolutionary history of living organisms is typically represented graphically by a *phylogeny*, a tree whose branchings indicate past speciation events. The inference of phylogenies based on molecular sequences extracted from extant species is a major task of computational biology. Among the many biological phenomena that complicate this task, one that has received much attention in the statistical phylogenetics literature is the variation in mutation rate across sites in a genome. (See related work below.) Such variation is generally attributed to unequal degrees of selective pressure. As we describe formally below, mathematically this phenomenon can be modeled as a *mixture* of phylogenies. That is, interpreting branch length as a measure of the amount of evolutionary change, rates-across-sites (RAS) models posit that all sites in a genome evolve according to a common tree topology, but branch lengths for a given site are scaled by a random factor.

Here we introduce a new approach for estimating RAS models. Our main contribution is a novel algorithm which *clusters* the sites according to their mutation rate. We show that our technique may be used to reconstruct phylogenies. Indeed, following the site clustering step, standard reconstruction techniques can be employed to recover a phylogeny on the unmixed subset of sites obtained. Our results rely on the following basic insight: there exist simple site-wise statistics that experience concentration-of-measure phenomena. Consequently, our techniques only hold *in the large-tree limit*.

Concentration has been used extensively in statistical phylogenetics. However its typical use is in the *large-sample limit*, that is, as the sequence length grows to infinity, for instance in order to show that so-called evolutionary distance estimates are accurate given sufficiently long sequences (see e.g. [ESSW99]). Instead, we consider here concentration in what we call the *large-tree limit*, that is, as the number of leaves goes to infinity. Note that the latter is trickier to analyze. Indeed, whereas different sites are usually assumed to evolve independently, leaf states are *not* independent. To the best of our knowledge, this is the first use of this type of concentration in the context of phylogenetics.

Our results imply, in particular, that large phylogenies are typically identifiable under rate variation. We also derive sequence-length requirements for high-probability reconstruction.

1.1 Related work

Most prior theoretical work on mixture models has focused on the question of *identifiability*. A class of phylogenetic models is identifiable if any two models in the class produce different data distributions. It is well-known that unmixed phylogenetic models are typically identifiable [Cha96]. *This is not the case in general for mixtures of phylogenies.* For instance, Steel et al. [SSH94] showed that for any two trees one can find a random scaling on each of them such that their data distributions are identical. Hence it is hopeless in general to reconstruct phylogenies under mixture models. See also [EW04, MS07, MMS08, SV07b, ŠV07a, Ste09] for further examples of this type.

However the negative examples constructed in the references above are not necessarily typical. They use special features of the mutation models (and their invariants) and allow themselves quite a bit of flexibility in setting up the topologies and branch lengths. In fact, recently a variety of more standard mixture models have been shown to be identifiable. These include the common GTR+Gamma model [AAR08, WS10] and GTR+Gamma+I model [CH11], as well as some covarion models [AR06], some group-based models [APRS11], and so-called r-component identical tree mixtures [RS10]. Although these results do not provide practical algorithms for reconstructing the corresponding mixtures, they do give hope that these problems may be tackled successfully.

Beyond the identifiability question, there seems to have been little rigorous work on reconstructing phylogenetic mixture models. One positive result is the case of the molecular clock assumption with across-sites rate variation [SSH94], although no sequence-length requirements are provided. There is a large body of work on practical reconstruction algorithms for various types of mixtures, notably rates-across-sites models and covarion-type models, using mostly likelihood and bayesian methods. See e.g. [Fel04] for references. But the optimization problems they attempt to solve are likely NP-hard [CT06, Roc06]. There also exist many techniques for testing for the presence of a mixture (for example, for testing for rate heterogeneity), but such tests typically require the knowledge of the phylogeny. See e.g. [HR97].

Here we give both identifiability and reconstruction results. Whereas Steel et al. [SSH94] show that any two fixed trees can be made indistinguishable with an appropriate (arbitrarily complex) choice of scaling distributions, we show in essence that, given a fixed rate distribution (or a well-behaved class of rate distributions), sufficiently large trees are typically distinguishable. After a draft of our results were circulated [MR08], related results for large trees were established

by Rhodes and Sullivant [RS10] using different techniques. In particular, our technical assumptions are similar in spirit to the genericity condition in [RS10]. Although our genericity assumptions are stronger, they allow an efficient reconstruction of the model and explicit bounds on sequence-length requirements. Note moreover that our results apply to general, possibly continuous, nonparametric rate distributions.

The proof of our main results relies on the construction of a *site clustering statistic* that discriminates between different rates. A similar statistic was also used in [SS06] in a different context. However, in contrast to [SS06], our main reconstruction result requires that a site clustering statistic be constructed based only on data generated by the mixture—that is, *without* prior knowledge of the model.

1.2 Overview of techniques

A simplified setting To illustrate our main ideas, we first consider a simple two-speed model. Assume that molecular sequences have two types of sites: "slow" and "fast." Both types of sites evolve independently by single substitution on a common evolutionary tree according, say, to a standard Jukes-Cantor model of substitution, but the fast ones evolve three times as fast. See Section 2 for a formal definition of the Jukes-Cantor model. To keep things simple, assume for now that the evolutionary tree is a complete binary tree with $n = 2^h$ leaves, where h is the number of levels. (Note that our results apply to much more general rate distributions. We also discuss how to deal with general trees. See below.)

Our approach is based on the following question: Is it possible to tell *with high confidence* which sites are slow or fast, with no prior knowledge of the phylogeny that generated them? Perhaps surprisingly, the answer is *yes*—at least for large trees. This far-reaching observation does not seem to have been made previously.

To see how this works, assume for the time being that we know the phylogeny. We will show how to remove this assumption below. Take a pair of leaves a, b. The effect of the speed of a site can be seen in the probability of agreement between a and b: the leaves agree more often on slow-evolving sites. Hence, if a site shows agreement between a and b, one may deduce that the site is more likely to be slow-evolving. But this is too little information to infer *with high confidence* the speed of a site. Instead, one may look at a larger collection of pairs of leaves and consider the statistic that counts how many of them agree on a given site. The idea is that a large number of agreements should indicate a slow site. For this scheme to work accurately, we require two properties from this statistic: *separa-*

tion and *concentration*. By separation, we mean that the expected value of the statistic should be different on slow and fast sites—in order to distinguish them. By concentration, we mean that the statistic should lie very close to its expectation. These two properties produce a good site clustering test. To satisfy them, the pairs of leaves involved must be chosen carefully.

Separation and concentration To obtain separation, it is natural to use only pairs of "close" leaves. Indeed, leaves that are far away are practically independent and the speed of a site has very little noticeable effect on their agreement. As for concentration, what one needs is the kind of conditions that give rise to the central limit theorem: a large sum of small independent contributions. For symmetric models such as the Jukes-Cantor model, the *agreement events* on two pairs of leaves (a, b) and (c, d) are independent as long as the paths between (a, b) and (c, d) do not intersect. Therefore, we are led to consider the following statistic: count how many cherries (that is, sister leaves) agree and divide by the total number of cherries to obtain a fraction. One can show from the considerations above that such a statistic is highly concentrated.

Unknown, general tree However our derivation *so far* has relied heavily on two unsatisfied premises:

1. That the tree is known. This is of course not the case since our ultimate goal is precisely to reconstruct the phylogeny.

2. And that the tree is complete. In particular, our argument uses the fact that complete binary trees contain many cherries. But general trees may have very few cherries.

Perhaps surprisingly, neither of these conditions is necessary. The bulk of the technical contributions of this paper lie in getting rid of these assumptions. We show in particular how to construct a site clustering statistic similar to the one above directly from the data without prior knowledge of the tree. At a high level, all one needs is to select a large collection of "sufficiently correlated" pairs of leaves and then "dilute" them to discard pairs that are too close to each other. This leads to a highly concentrated site-wise statistic. See Section 2 for a statement of our results.

2 Definitions and Results

2.1 Basic Definitions

Phylogenies A phylogeny is a graphical representation of the speciation history of a group of organisms. The leaves typically correspond to current species. Each branching indicates a speciation event. Moreover we associate to each edge a positive weight. This weight can be thought roughly as the time elapsed on the edge multiplied by the mutation rate which may also depend on the edge. More formally:

Definition 1 (Phylogeny) *A* phylogeny $T = (V, E; L, \mu)$ *is a tree with vertex set* V*, edge set* E *and* n *(labelled) leaves* $L = [n] = \{1, \ldots, n\}$ *such that 1) the degree of all internal vertices* $V - L$ *is exactly* 3*, and 2) the edges are assigned weights* $\mu : E \to (0, +\infty)$*. We let* $\mathcal{T}[T] = (V, E; L)$ *be the* topology *of* T*. A phylogeny is naturally equipped with a so-called* tree metric *on the leaves* $d : L \times L \to (0, +\infty)$ *defined as follows*

$$\forall u, v \in L,\ d(u, v) = \sum_{e \in \mathrm{Path}_T(u,v)} \mu_e,$$

where $\mathrm{Path}_T(u, v)$ *is the set of edges on the path between* u *and* v *in* T*. We will refer to* $d(u, v)$ *as the* evolutionary distance *between* u *and* v*. Since under the assumptions above there is a one-to-one correspondence between* d *and* μ *(see e.g. [SS03]), we write either* $T = (V, E; L, d)$ *or* $T = (V, E; L, \mu)$*. We also sometimes use the natural extension of* d *to the internal vertices of* T*.*

We will sometimes restrict ourselves to the following standard special case.

Definition 2 (Regular Phylogenies) *Let* $0 < f \leq g < +\infty$*. We denote by* $\mathcal{T}_{f,g}$ *the set of phylogenies* $T = (V, E; L, \mu)$ *such that* $\forall e \in E,\ f \leq \mu_e \leq g$*.*

Poisson Model A standard model of DNA sequence evolution is the following *Poisson model*. See e.g. [SS03].

Definition 3 (Poisson Model) *Consider the following stochastic process. We are given a phylogeny* $T = (V, E; [n], \mu)$ *and a finite set* $\mathcal{R}$ *with* r *elements. Let* π *be a probability distribution on* $\mathcal{R}$*. Let* $Q \in \mathbb{R}^{r \times r}$ *be the following* rate matrix

$$Q_{xy} = \begin{cases} \pi_y, & \text{if } x \neq y, \\ \pi_y - 1, & \text{o.w.} \end{cases}$$

Associate to each edge $e \in E$ the stochastic matrix

$$[M(e)]_{xy} = [\exp(\mu_e Q)]_{xy} = \begin{cases} \pi_x + (1-\pi_x)e^{-\mu_e}, & \text{if } x = y, \\ \pi_y(1-e^{-\mu_e}), & \text{o.w.} \end{cases}$$

The process runs as follows. Choose an arbitrary root $\rho \in V$. Denote by $E_\downarrow$ the set E directed away from the root. Pick a state for the root according to π. Moving away from the root toward the leaves, apply the channel $M(e)$ to each edge e independently. Denote the state so obtained $\sigma_V = (\sigma_v)_{v \in V}$. In particular, $\sigma_{[n]}$ is the state at the leaves. More precisely, the joint distribution of σ_V is given by

$$\mu_V(\sigma_V) = \pi_\rho(\sigma_\rho) \prod_{e=(u,v) \in E_\downarrow} [M(e)]_{\sigma_u \sigma_v}.$$

For $W \subseteq V$, we denote by μ_W the marginal of μ_V at W. Under this model, the weight μ_e is the expected number of substitutions on edge e in a related continuous-time process. The r-state Poisson model *is the special case when π is the uniform distribution over $\mathcal{R}$. In that case, we denote the distribution of σ_V by $\mathcal{D}[T, r]$. When r is clear from the context, we write instead $\sigma_V \sim \mathcal{D}[T]$.*

More generally, we take k independent samples $(\sigma_V^i)_{i=1}^k$ from the model above, that is, $\sigma_V^1, \ldots, \sigma_V^k$ are i.i.d. $\mathcal{D}[T, r]$. We think of $(\sigma_v^i)_{i=1}^k$ as the sequence at node $v \in V$. Typically, $\mathcal{R} = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$ and the model describes how DNA sequences stochastically evolve by point mutations along an evolutionary tree—under the assumption that each site in the sequences evolves independently.

Example 1 (CFN and Jukes-Cantor Models) *The special case $r = 2$ corresponds to the so-called CFN model. The special case $r = 4$ is the well-known Jukes-Cantor model.*

We fix r throughout.

Remark 1 *We discuss the more general GTR model in the concluding remarks.*

Rates-across-sites model We introduce the basic rates-across-sites model which will be the focus of this paper. We will use the following definition.

Definition 4 (Phylogenetic Scaling) *Let $T = (V, E; [n], \mu)$ be a phylogeny and Λ, a constant in $[0, +\infty)$. Then we denote by ΛT the phylogeny obtained by scaling the weights of T by Λ, that is, $\Lambda T = (V, E; [n], \Lambda\mu)$.*

Definition 5 (Rates-Across-Sites Model (see e.g. [SSH94])) *In the* generalized Poisson model *we are given a phylogeny T and a* scaling factor Λ*, that is, a random variable on $[0, +\infty)$. Let $\Lambda_1, \ldots, \Lambda_k$ be i.i.d. copies of Λ. Conditioned on $\Lambda_1, \ldots, \Lambda_k$, the samples $(\sigma_V^i)_{i=1}^k$ generated under this model are independent with $\sigma_V^j \sim \mathcal{D}[\Lambda_j T]$, $j = 1, \ldots, k$. We denote by $\mathcal{D}[T, \Lambda, r]$ the probability distribution of σ_V^1. We also let $\overline{\mathcal{D}}[T, \Lambda, r]$ be the probability distribution of σ_L^1.*

2.2 Main results

Tree identifiability To provide a *uniform* bound on the minimum tree size required for our identifiability result to hold, we make explicit assumptions on the mutation model. For $s \geq 0$, let

$$\Phi(s) = \mathbb{E}\left[e^{-s\Lambda}\right],$$

be the *moment generating function* (or one-sided Laplace transform) of the scaling factor Λ. The probability distribution of Λ is determined by Φ. See e.g. [Bil95]. We normalize Λ so that

$$-\Phi'(0) = \mathbb{E}[\Lambda] = 1.$$

In particular, Λ is not identically 0 and Φ is continuous and strictly decreasing.

Assumption 1 *Let $0 < f \leq g < +\infty$, and $M > 0$. The following set of assumptions on a generalized Poisson model will be denoted by $\mathrm{A}(f, g, M)$:*

1. Regular phylogeny: *The phylogeny $T = (V, E; [n], \mu)$ is in $\mathcal{T}_{f,g}$.*
2. Mass close to 0: *We have that*

$$\Phi^{-1}\left(e^{-6g}\right) \leq M.$$

In words, an evolutionary distance of M under Λ-scaling produces a correlation corresponding to an evolutionary distance of at least $6g$ without the scaling. We denote by $\mathrm{GPM}(f, g, M, n_0)$ the set of generalized Poisson models satisfying $\mathrm{A}(f, g, M)$ with at least n_0 leaves.

Remark 2 *Note that the results in [SSH94] indicate that appropriate conditions are needed to obtain a tree identifiability result in the generalized Poisson model when the random scaling is unknown. We do not claim that the conditions above are minimal. The first assumption is meant to ensure that there is enough signal to*

reconstruct the tree. The second assumption bounds the distortion of the random scaling for evolutionary distances corresponding to short paths. It essentially implies that the probability mass of Λ close to 0 is bounded. In particular, note that if the probability mass below

$$\varepsilon = -\frac{1}{M} \ln \left(\frac{e^{-6g} - \delta}{1 - \delta} \right),$$

is less than δ (for $\delta < e^{-6g}$), then

$$\Phi(M) \leq \delta + (1 - \delta) e^{-\varepsilon M} \leq e^{-6g},$$

and the assumption is satisfied. Conversely, if Λ satisfies the second assumption then the probability mass δ below ε (for $\varepsilon < 6g/M$) must be such that

$$\delta \leq e^{-(6g - \varepsilon M)},$$

since

$$e^{-6g} \geq \Phi(M) \geq \delta e^{-\varepsilon M}.$$

Remark 3 *The second assumption implicitly implies that*

$$\mathbb{P}[\Lambda = 0] \leq e^{-6g}.$$

This is in fact not necessary. By first removing all invariant sites, it should be possible to extend our main theorem to moment generating functions of the form

$$\Phi(s) = \alpha + (1 - \alpha) \Phi_+(s),$$

where $0 \leq \alpha < 1$ is uniformly bounded away from 1 and Φ_+ satisfies the assumptions above. Indeed, on a large phylogeny, it is extremely unlikely to produce an invariant site using a positive scaling factor. Hence removing all invariant sites has the effect of essentially restricting the dataset to the positive part of the distribution of Λ. We leave the details to the reader. Given this observation, in the rest of the manuscript one can assume that

$$\mathbb{P}[\Lambda = 0] = 0.$$

Theorem 1 (Tree identifiability) *Fix $0 < f \leq g < +\infty$, and $M > 0$. Then, there exists $n_0(f, g, M) \geq 1$ such that, if (T, Λ) and (T', Λ') are in $\mathrm{GPM}(f, g, M, n_0)$ with $\mathcal{T}[T] \neq \mathcal{T}[T']$, then*

$$\overline{\mathcal{D}}[T, \Lambda, r] \neq \overline{\mathcal{D}}[T', \Lambda', r].$$

(Recall that $\overline{\mathcal{D}}[T, \Lambda, r]$ denotes the distribution at the leaves*.)*

Remark 4 *Note that we allow $\Lambda \sim \Lambda'$ (where $\sim$ denotes equality in distribution). This is the sense in which our result is a* tree *identifiability result.*

Remark 5 *Note that our identifiability result applies only to* sufficiently large *phylogenies. Computing n_0 from our techniques is difficult (and in general depends on the parameters f, g, M). One could estimate the required size by running the reconstruction algorithm below on simulated data for various sizes and parameters. We leave such empirical studies for future work.*

The proof of our main theorem relies on the following reconstruction result.

Tree reconstruction Moreover, we give a stronger result implying that the phylogeny can be reconstructed with high confidence using polynomial length sequences in polynomial time. The proof appears in Sections 3, 4 and 5.

Theorem 2 (Tree Reconstruction) *Under Assumption 1, for all $0 < \delta < 1$, there is a $\gamma_k > 0$ large enough so that the topology of the tree can be reconstructed in polynomial time using $k = n^{\gamma_k}$ samples, except with probability δ.*

Remark 6 *Once the tree has been estimated, one can also infer the rate distribution. Details are left to the interested reader.*

3 Site clustering statistic: Existence and properties

In this section, we introduce our main site clustering statistic and show that it is concentrated. Let $0 < f \leq g < +\infty$ and $M > 0$. In this section, we fix a phylogeny $T = (V, E; [n], \mu)$ in $\mathcal{T}_{f,g}$. We let $(\sigma_L^i)_{i=1}^k$ be k i.i.d. samples from $\overline{\mathcal{D}}[T, \Lambda, r]$ where the generalized Poisson model (T, Λ) satisfies Assumption 1. Moreover, let $\Lambda_1, \ldots, \Lambda_k$ be the i.i.d. scaling factors corresponding to the k samples above.

Some notation We will also use the notation $[n]^2 = \{(a, b) \in [n] \times [n] \,:\, a \leq b\}$, $[n]^2_= = \{(a, a)\}_{a \in [n]}$, and $[n]^2_{\neq} = [n]^2 - [n]^2_=$. We also denote by $[n]^4_{\neq}$ the set of pairs $(a, b), (c, d) \in [n]^2_{\neq}$ such that $(a, b) \neq (c, d)$ (as pairs). For $\alpha > 0$, we let

$$\Upsilon_\alpha = \{(a, b) \in [n]^2_{\neq} \,:\, d(a, b) \leq \alpha\},$$

be all pairs of leaves in T at evolutionary distance at most α. Let $p_\infty = 1 - q_\infty$ where $q_\infty = \sum_{x \in \mathcal{R}} \pi_x^2$.

3.1 What makes a good site clustering statistic?

For a site $i = 1, \ldots, k$, consider a statistic of the form

$$\mathcal{U}_i = \frac{1}{|\Upsilon|} \sum_{(a,b)\in\Upsilon} p_\infty^{-1}[\mathbb{1}\{\sigma_a^i = \sigma_b^i\} - q_\infty], \tag{1}$$

where $\Upsilon \subseteq [n]^2_{\neq}$, is a subset of pairs of distinct leaves independent of i. Using the expression for the transition matrix given in Definition 3, note that

$$\begin{aligned}
\mathbb{E}[\mathcal{U}_i \mid \Lambda_i] &= \frac{1}{|\Upsilon|} \sum_{(a,b)\in\Upsilon} p_\infty^{-1} \left[\mathbb{E}[\mathbb{1}\{\sigma_a^i = \sigma_b^i\} \mid \Lambda_i] - q_\infty\right] \\
&= \frac{1}{|\Upsilon|} \sum_{(a,b)\in\Upsilon} p_\infty^{-1} \left[\sum_{x\in\mathcal{R}} \pi_x \left(\pi_x + (1-\pi_x)e^{-\Lambda_i d(a,b)}\right) - q_\infty\right] \\
&= \frac{1}{|\Upsilon|} \sum_{(a,b)\in\Upsilon} e^{-\Lambda_i d(a,b)},
\end{aligned}$$

which is *strictly decreasing in* Λ_i. We need two properties for (1) to make a good site clustering statistic: separation and concentration.

For *separation*, that is, for the statistic above to distinguish different scaling factors as much as possible, we require the following condition:

S1 Each pair in Υ is composed of two "sufficiently close" leaves, that is, there is $\alpha < +\infty$ such that $\Upsilon \subseteq \Upsilon_\alpha$.

Indeed, if two leaves are far away, their joint distribution is close to independent and scaling has little effect on their agreement. A much better separation is obtained from close leaves.

To guarantee *concentration* of a statistic of the type (1), we require the following three conditions on Υ:

C1 The set Υ is "large enough" and each pair makes a "small contribution" to the sum. This will be satisfied if we show that we can take $|\Upsilon| = \Theta(n)$, as (1) is a sum of $\{0,1\}$-variables.

C2 Agreement for different pairs in Υ is "sufficiently uncorrelated," e.g., independent.

These conditions will allow us to apply standard large deviations arguments.

Example 2 (Full sum) *As a first guess, one may expect that taking* Υ *to be* all pairs of leaves *may give a good site clustering statistic. However, in general this is not the case as we show in the following example. Consider the two-state case on a complete binary tree with identical edge lengths* μ *and* $\Lambda = 1$*. For mathematical convenience, assume that the states are* $\mathcal{R} = \{+1, -1\}$ *and let*

$$\gamma = 2e^{-2\mu}.$$

Then, up to a multiplicative factor and additive constant, the clustering statistic is simply

$$\mathcal{U}^{(h)} = \sum_{(a,b)\in[n]^2_{\neq}} \sigma_a\sigma_b,$$

for a tree with h *levels. Using a calculation of [EKPS00, Section 5], one has*

$$\mathbb{E}[\sigma_a\sigma_b] = e^{-d(a,b)}. \tag{2}$$

Dividing the expectation into terms over the first subtree of the root, terms over the second subtree of the root, and terms between the two subtrees, we have

$$\mathbb{E}\left[\mathcal{U}^{(h)}\right] = 2\mathbb{E}\left[\mathcal{U}^{(h-1)}\right] + (2^{h-1})^2 e^{-2h\mu}.$$

Solving for the recursion gives

$$\mathbb{E}\left[\mathcal{U}^{(h)}\right] = \gamma 2^{h-2}\frac{\gamma^h - 1}{\gamma - 1} = O\left(2^{2h}\gamma^{2h}\right),$$

as $h \to \infty$*. On the other hand, the expectation of the square* $\mathbb{E}[(\mathcal{U}^{(h)})^2]$ *is a sum of terms of the form* $\mathbb{E}[\sigma_{z_1}\sigma_{z_2}\sigma_{z_3}\sigma_{z_4}]$ *where some of the* z*'s may be repeated. All such terms are non-negative because of* (2) *and the fact that terms where all* z*'s are different factor into a product by Proposition 1 below. Hence*

$$\begin{aligned}
\mathrm{Var}\left[\mathcal{U}^{(h)}\right] &\geq \sum_{(a,b)\in[n]^2_{\neq}} \mathbb{E}\left[(\sigma_a\sigma_b)^2\right] - \mathbb{E}\left[\mathcal{U}^{(h)}\right]^2 \\
&= \frac{2^h(2^h-1)}{2} - \gamma^2 2^{2h-4}\left(\frac{\gamma^h - 1}{\gamma - 1}\right)^2 \\
&= \Omega(2^{2h}),
\end{aligned}$$

if $\gamma < 1$*. Hence, assuming that* $\gamma < 1$*, we have*

$$\frac{\mathbb{E}\left[\mathcal{U}^{(h)}\right]^2}{\mathrm{Var}\left[\mathcal{U}^{(h)}\right]} \to 0,$$

as $h \to \infty$*. In other words, the sum over all pairs is too "noisy" to serve as a site clustering statistic in that case.*

3.2 Does it exist?

We now show that there always exist statistics that satisfy the properties above and we give explicit guarantee on their concentration. Note that, in the current section, we only provide a proof of *existence*. In particular, in establishing existence, we use evolutionary distances which are not available from the data. Later, in Section 4, we explain how to *construct* such a statistic from the data $\overline{\mathcal{D}}[T, \Lambda, r]$ (or, more precisely, the samples $(\sigma_L^i)_{i=1}^k$) *without knowledge of the tree topology, evolutionary distances, or site scaling factors*.

We now explain how the conditions above can be achieved for an appropriate choice of Υ on *any* tree topology. Note, however, that Υ depends on T.

Independence We first show that the clustering statistic (1) is a sum of independent variables *as long as the paths between different pairs do not intersect.* This will allow us to satisfy C2.

Proposition 1 (Independence) *Assume that for all* $(a, b), (a', b') \in \Upsilon$ *with* $(a, b) \neq (a', b')$ *we have*

$$\mathrm{Path}(a, b) \cap \mathrm{Path}(a', b') = \emptyset,$$

where $\mathrm{Path}(a, b)$ *is the set of edges on the path between* a *and* b*. Then the random variables* $\{\mathbb{1}\{\sigma_a = \sigma_b\}\}_{(a,b)\in\Upsilon}$*, are mutually independent.*

Proof: Denote $\Upsilon = \{(a_1, b_1), \ldots, (a_\upsilon, b_\upsilon)\}$ with $\upsilon = |\Upsilon|$. Let $\mathcal{V}$ be the set of nodes on the path between a_1 and b_1. Removing the edges in $\mathrm{Path}(a_1, b_1)$ creates a forest where the $\mathcal{V}$-nodes can be taken as roots. Note that, by symmetry and the Markov property, conditioned on $\mathcal{V}$, the distribution of the random variables

$$\left\{\mathbb{1}\{\sigma_a = \sigma_b\}\right\}_{(a,b)\in\Upsilon-\{(a_1,b_1)\}}, \tag{3}$$

does not depend on the states of the $\mathcal{V}$-nodes. In particular, $\mathbb{1}\{\sigma_{a_1} = \sigma_{b_1}\}$ is independent of (3). Proceeding by induction gives the result. ■

Size To satisfy S1, we restrict ourselves to "close pairs." We first show that the size of Υ_α grows linearly *as long as* $\alpha \geq 4g$, allowing us to also satisfy C1. A similar result is proved in [SS06].

Proposition 2 (Size of Υ_α) *Let* $\alpha \geq 4g$*. Then*

$$|\Upsilon_\alpha| \geq \frac{n}{4}.$$

Proof: Let

$$\Gamma = \{a \in [n] \ : \ d(a,b) > \alpha, \ \forall b \in [n] - \{a\}\},$$

that is, Γ is the set of leaves with no other leaf at evolutionary distance α. We will bound the size of Γ. For $a \in \Gamma$, let

$$\mathcal{B}(a) = \left\{v \in V \ : \ d(a,v) \leq \frac{\alpha}{2}\right\}.$$

Note that for all $a, b \in \Gamma$ with $a \neq b$ we have $\mathcal{B}(a) \cap \mathcal{B}(b) = \emptyset$ by the triangle inequality. Moreover, it holds that for all $a \in \Gamma$

$$|\mathcal{B}(a)| \geq 2^{\left\lfloor \frac{\alpha}{2g} \right\rfloor},$$

since T is binary and there is no leaf other than a in $\mathcal{B}(a)$. Hence, we must have

$$|\Gamma| \leq \frac{2n-2}{2^{\left\lfloor \frac{\alpha}{2g} \right\rfloor}} \leq \left(\frac{1}{2^{\left\lfloor \frac{\alpha}{2g} \right\rfloor - 1}}\right) n,$$

as there are $2n-2$ nodes in T.

Now, for all $a \notin \Gamma$ assign an arbitrary leaf at evolutionary distance at most α. Then

$$\begin{aligned} |\Upsilon_\alpha| &\geq \frac{1}{2}(n - |\Gamma|) \\ &\geq \frac{1}{2}\left(1 - \frac{1}{2^{\left\lfloor \frac{\alpha}{2g} \right\rfloor - 1}}\right) n, \end{aligned}$$

where we divided by 2 to avoid over-counting. The result follows from the assumption $\alpha \geq 4g$. ■

Sparsification Note that Υ_{4g} satisfies C1 but does not satisfy C2 as the pairs may be intersecting (see Proposition 1). We now show how to satisfy both C1 and C2 by "sparsifying" Υ_{4g}. In stating this procedure, we allow some flexibility (that is, arbitrary choices) which will be useful in analyzing the actual implementation in the next section. Let $4g < m < M$ be a constant to be determined later and assume Υ' is any set satisfying

$$\Upsilon_{4g} \subseteq \Upsilon' \subseteq \Upsilon_m.$$

We know from Proposition 2 that Υ' has linear size, that is, $|\Upsilon'| \geq n/4$. We construct a linear-sized subset Υ of Υ' satisfying the non-intersection condition of Proposition 1 as follows. Let $S := \Upsilon'$ and $\Upsilon'' := \emptyset$.

- Take any pair (a^*, b^*) in S and add it to Υ''.
- Let S_0 be any subset of S such that S_0 contains all pairs with at least one node within evolutionary distance m of either a^* or b^* and contains no pair with both nodes beyond evolutionary distance M from both a^* and b^*. Remove S_0 from S.
- Repeat until S is empty.
- Return $\Upsilon := \Upsilon''$.

We claim that Υ is linear in size and that no two pairs in Υ intersect.

Proposition 3 (Properties of Υ) *Let Υ be any set built by the procedure above. Then,*

1. *For all $(a, b), (a', b') \in \Upsilon$ with $(a, b) \neq (a', b')$ we have $\mathrm{Path}(a, b) \cap \mathrm{Path}(a', b') = \emptyset$.*
2. *There is $\gamma_s = \gamma_s(M, f) > 0$ such that $|\Upsilon| \geq \gamma_s n$, where γ_s does not depend on T, but only on M, f.*
3. *For all $(a, b) \in \Upsilon$, we have*

$$2f \leq d(a, b) \leq M.$$

Proof: We first prove the non-intersecting condition. All pairs of leaves in Υ are at evolutionary distance at most m. Moreover, for any $(a, b) \neq (a', b')$ in Υ, we have by construction

$$\min\{d(u, v) : u \in \{a, b\}, v \in \{a', b'\}\} \geq m.$$

Hence, the path between a and b and the path between a' and b' cannot intersect: we have

$$d(a, b) + d(a', b') - d(a, a') - d(b, b') \leq 0,$$

which, using $\mu_e > 0$ for all e and the four-point test (see e.g. [SS03]), excludes the topology $aa'|bb'$ (that is, the four-leaf topology where $\{a, a'\}$ is one side of the internal edge and $\{b, b'\}$ is on the other); and similarly for the topology $ab'|a'b$.

We now bound the size of Υ. Let (a, b) be a pair of leaves at evolutionary distance at most m. There are at most $2 \cdot 2^{\lfloor \frac{M}{f} \rfloor - 1} = 2^{\lfloor \frac{M}{f} \rfloor}$ leaves at evolutionary distance at most M from either a or b. Therefore, at each iteration of

the sparsification algorithm, the number of elements of S removed is at most $2^{\lfloor \frac{M}{f} \rfloor + \lfloor \frac{m}{f} \rfloor - 1} \leq 2^{2\lfloor \frac{M}{f} \rfloor}$, as each leaf removed is involved in at most $2^{\lfloor \frac{m}{f} \rfloor - 1}$ pairs at evolutionary distance m. Since by Proposition 2, the size of Υ' is at least $n/4$, the number of elements in Υ at the end of the sparsification algorithm is at least

$$|\Upsilon| \geq \frac{n}{2^{2\lfloor \frac{M}{f} \rfloor + 2}}.$$

Finally, by our assumption on the phylogeny, two distinct leaves are always at evolutionary distance at least $2f$. For the upper bound, use $m < M$. ■

Define

$$\gamma_s = \gamma_s(M, f) = \frac{1}{2^{2\lfloor \frac{M}{f} \rfloor + 2}}.$$

Definition 6 (Sparse pairs) *We say that a set $\Upsilon \subseteq [n]^2_{\neq}$ is* γ_s-sparse *if it satisfies the three properties in the statement of Proposition 3.*

3.3 Properties of the site clustering statistic

In (1) fix an γ_s-sparse Υ. We now show that conditions C1 and C2 lead to concentration. Let

$$U_\Upsilon = \mathbb{E}[\mathcal{U}_i] = \frac{1}{|\Upsilon|} \sum_{(a,b)\in\Upsilon} \mathbb{E}\left[e^{-\Lambda_i d(a,b)}\right] = \frac{1}{|\Upsilon|} \sum_{(a,b)\in\Upsilon} \Phi(d(a,b)).$$

Moreover, for $\lambda \geq 0$, define

$$U_\Upsilon(\lambda) = \frac{1}{|\Upsilon|} \sum_{(a,b)\in\Upsilon} e^{-\lambda d(a,b)}$$

and note that

$$U_\Upsilon(\Lambda_i) = \mathbb{E}[\mathcal{U}_i \,|\, \Lambda_i],$$

and

$$U_\Upsilon = \mathbb{E}\left[U_\Upsilon(\Lambda_i)\right],$$

for $i = 1, \ldots, k$.

Proposition 4 (Concentration of $\mathcal{U}_i$) *For all $\zeta > 0$, there is $c > 0$ depending on M, f such that*

$$\mathbb{P}\left[|\mathcal{U}_i - U_\Upsilon(\Lambda_i)| \geq \zeta \,|\, \Lambda_i\right] \leq 2\exp(-c\zeta^2 n),$$

almost surely, for all $i = 1, \ldots, k$. (We will eventually use $\zeta = o(1)$.)

Proof: Recall the following standard concentration inequality (see e.g. [MR95]):

Lemma 1 (Azuma-Hoeffding Inequality) *Suppose $X = (X_1, \ldots, X_m)$ are independent random variables taking values in a set S, and $f : S^m \to \mathbb{R}$ is any t-Lipschitz function: $|f(\mathbf{x}) - f(\mathbf{y})| \leq t$ whenever $\mathbf{x}$ and $\mathbf{y}$ differ at just one coordinate. Then, $\forall \zeta > 0$,*

$$\mathbb{P}\left[|f(X) - \mathbb{E}[f(X)]| \geq \zeta\right] \leq 2\exp\left(-\frac{\zeta^2}{2t^2 m}\right).$$

From Propositions 1, 2, and 3, the random variable $\mathcal{U}_i$ is a (normalized) sum of $\Omega(n)$ independent bounded variables. By Lemma 1, conditioning on Λ_i, we have $|U_\Upsilon(\Lambda_i) - \mathcal{U}_i| \leq \zeta$ except with probability $\exp(-\Omega(\zeta^2 n))$, where we used that $m = \Omega(n)$ and $t = O(1/n)$. ■

Moreover, we show separation.

Proposition 5 (Separation of $\mathcal{U}_i$) *If $\lambda - \lambda' \geq \beta$, where $\beta \geq 0$, then*

$$U_\Upsilon(\lambda') - U_\Upsilon(\lambda) \geq e^{-\lambda M}\left(e^{2f\beta} - 1\right).$$

Proof: We have

$$\begin{aligned}
U_\Upsilon(\lambda') - U_\Upsilon(\lambda) &= \frac{1}{|\Upsilon|}\sum_{(a,b)\in\Upsilon}\left[e^{-\lambda' d(a,b)} - e^{-\lambda d(a,b)}\right] \\
&\geq \frac{1}{|\Upsilon|}\sum_{(a,b)\in\Upsilon}\left[e^{-(\lambda-\beta) d(a,b)} - e^{-\lambda d(a,b)}\right] \\
&\geq \frac{1}{|\Upsilon|}\sum_{(a,b)\in\Upsilon} e^{-\lambda d(a,b)}\left[e^{\beta d(a,b)} - 1\right] \\
&\geq \frac{1}{|\Upsilon|}\sum_{(a,b)\in\Upsilon} e^{-\lambda M}\left[e^{\beta\cdot 2f} - 1\right],
\end{aligned}$$

since $2f \leq d(a,b) \leq M$ for all $(a,b) \in \Upsilon$ by assumption. ■

Remark 7 *The last bound may seem problematic because under our assumptions the scaling factor is allowed to have an unbounded support. In that case the RHS could be arbitrarily close to 0. But we will show below that we can safely ignore large values of λ.*

4 Constructing the site clustering statistic from data

Note that in the previous section we only established the *existence* of an appropriate site clustering statistic. We now show how such a statistic can be built from data *without knowledge of the tree topology or site scaling factors*.

Notation We use the same notation as in the previous section. Further, we let

$$\hat{q}(a,b) = \frac{1}{k}\sum_{i=1}^{k} p_\infty^{-1}\left[\mathbb{1}\{\sigma_a^i = \sigma_b^i\} - q_\infty\right].$$

Also let

$$\begin{aligned} q(a,b) &= \mathbb{E}[\hat{q}(a,b)] \\ &= \mathbb{E}\left[p_\infty^{-1}\left[\mathbb{1}\{\sigma_a^1 = \sigma_b^1\} - q_\infty\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[p_\infty^{-1}\left[\mathbb{1}\{\sigma_a^1 = \sigma_b^1\} - q_\infty\right] \mid \Lambda_1\right]\right] \\ &= \mathbb{E}\left[e^{-\Lambda_1 d(a,b)}\right] \\ &= \Phi(d(a,b)). \end{aligned}$$

where we used our previous calculations. We define some constants used in the algorithm and its analysis whose values will be justified below. Let

$$\omega_m = e^{-5g}, \qquad \omega_m^+ = e^{-5.5g}, \qquad \omega_m^- = e^{-4.5g}.$$

Also recall that Φ is strictly decreasing and let

$$m = \Phi^{-1}(\omega_m).$$

Note that by Jensen's inequality and $\mathbb{E}[\Lambda_1] = 1$

$$\Phi(5g) \geq e^{-\mathbb{E}[\Lambda_1]\cdot 5g} = e^{-5g},$$

so that

$$m \geq 5g > 4g,$$

as assumed in the previous section. Similarly, by assumption,

$$m = \Phi^{-1}(5g) < \Phi^{-1}(e^{-6g}) \leq M,$$

so that $m < M$. Finally let

$$\eta = \min\left\{e^{-4g} - e^{-4.5g}, e^{-4.5g} - e^{-5g}, e^{-5g} - e^{-5.5g}, e^{-5.5g} - e^{-6g}\right\}.$$

4.1 Site clustering algorithm

We proceed in three steps, as in the idealized setting of Section 3.2. However, unlike the idealized setting, we do *not* assume the knowledge of evolutionary distances. Note in particular that it is not possible to estimate $d(a, b)$ from the samples $(\sigma_L^i)_{i=1}^k$ because the rate distribution is unknown. Instead, we use $\hat{q}(a, b)$ as a *rough* estimate of how close a and b are in the tree. This will suffice for our purposes, as we show in the next subsection. The algorithm is the following:

1. *(Close Pairs)* For all pairs of leaves $a, b \in [n]$, compute $\hat{q}(a, b)$ and set

$$\Upsilon' = \{(a, b) \in [n]^2_{\neq} \ : \ \hat{q}(a, b) \geq \omega_m^-\}.$$

2. *(Sparsification)* Let $S := \Upsilon'$ and $\Upsilon'' := \emptyset$.

   - Take any pair (a^*, b^*) in S and add it to Υ''.
   - Remove from S all pairs (a, b) such that

$$\max\{\hat{q}(c^*, c) \ : \ c^* \in \{a^*, b^*\}, c \in \{a, b\}\} \geq \omega_m^+.$$

   - Repeat until S is empty.

3. *(Final Statistic)* Return $\Upsilon := \Upsilon''$.

4.2 Analysis of the clustering algorithm

Let Υ be the set returned by the previous algorithm. We show that it is γ_s-sparse with high probability.

Proposition 6 (Clustering statistic) *Under Assumption 1, for all $0 < \delta < 1$ there exists a constant $0 < C < +\infty$ (depending on g) such that the set of pairs Υ returned by the previous algorithm is γ_s-sparse with probability $1 - \delta$ provided that the number of samples satisfies*

$$k \geq C \log n.$$

Moreover, the algorithm runs in polynomial time.

Proof: We first prove that all $\hat{q}(a, b)$'s are sufficiently accurate.

Lemma 2 *For all $0 < \delta < 1$, there exists a constant $0 < C < +\infty$ (depending on g) such that*

$$|\hat{q}(a,b) - q(a,b)| \leq \eta,$$

for all $(a,b) \in [n]^2_{\neq}$ with probability $1-\delta$ provided that the number of samples satisfies

$$k \geq C \log n.$$

Proof: For each $(a,b) \in [n]^2_{\neq}$, $\hat{q}(a,b)$ is a sum of k independent bounded variables. By Lemma 1, taking $\zeta = \eta$ we have

$$|\hat{q}(a,b) - q(a,b)| \leq \eta,$$

except with probability $2\exp(-C'k)$ for some $C' > 0$ depending on p_∞ and η. Note that there are at most n^2 elements in $[n]^2_{\neq}$ so that the probability of failure is at most

$$2n^2 \exp(-C' \cdot C \log n) \leq \delta,$$

for C sufficiently large. ■

We return to the proof of Proposition 6. Assume that the conclusion of the previous lemma holds. Our goal is to prove that the site clustering algorithm then follows the idealized sparsification procedure described in Section 3.

1. We first prove that the set

   $$\Upsilon' = \{(a,b) \in [n]^2_{\neq} \ : \ \hat{q}(a,b) \geq \omega^-_m\}.$$

   satisfies

   $$\Upsilon_{4g} \subseteq \Upsilon' \subseteq \Upsilon_m.$$

   Let (a,b) be such that $d(a,b) \leq 4g$. Then

   $$q(a,b) = \Phi(d(a,b)) \geq \Phi(4g) \geq e^{-4g},$$

   by monotonicity and Jensen's inequality. Hence

   $$\hat{q}(a,b) \geq e^{-4g} - \eta \geq e^{-4g} - \left(e^{-4g} - e^{-4.5g}\right) \geq e^{-4.5g} = \omega^-_m,$$

   and $\Upsilon_{4g} \subseteq \Upsilon'$.

   Similarly, let (a,b) be such that $d(a,b) > m$. Then

   $$q(a,b) = \Phi(d(a,b)) < \Phi(m) = \omega_m,$$

   and

   $$\hat{q}(a,b) < \omega_m + \eta \leq e^{-5g} + \left(e^{-4.5g} - e^{-5g}\right) = \omega^-_m,$$

   so that $\Upsilon' \subseteq \Upsilon_m$.

2. Let Υ'' be the set obtained during one of the iterations of Step 2 of the site clustering algorithm and fix a pair $(a^*, b^*) \in \Upsilon''$. We need to show that the set S_0 of pairs (a, b) in Υ'' such that

$$\max\{\hat{q}(c^*, c) \ : \ c^* \in \{a^*, b^*\}, c \in \{a, b\}\} \geq \omega_m^+ \tag{4}$$

is such that it contains all pairs with at least one node within evolutionary distance m of either a^* or b^* and contains no pair with both nodes beyond evolutionary distance M from both a^* and b^*. In the first case, assume w.l.o.g. that

$$d(a, a^*) \leq m.$$

Then, arguing as above,

$$\hat{q}(a, a^*) \geq e^{-5g} - \left(e^{-5g} - e^{-5.5g}\right) = \omega_m^+,$$

and (4) is satisfied. In the second case, for all $c \in \{a, b\}$ and $c^* \in \{a^*, b^*\}$

$$d(c, c^*) > M,$$

and

$$\hat{q}(c, c^*) < e^{-6g} + \left(e^{-5.5g} - e^{-6g}\right) = \omega_m^+,$$

so that (4) is not satisfied.

The two properties above guarantee that the algorithm constructs a set Υ as in the idealized sparsification procedure of Section 3. In particular, Υ is γ_s-sparse by Proposition 3. ■

5 Tree reconstruction

We now show how to use our site clustering statistic to build the tree itself. The algorithm is composed of two steps: we first "bin" the sites according to the value of the clustering statistic; we then use the sites in one of those bins and apply a standard distance-based reconstruction method. By taking the bins sufficiently small, we show that the content of the bins is made of sites with almost identical scaling factor—thus essentially reducing the situation to the unmixed case.

Throughout this section, we assume that there is a $\gamma_k > 0$ such that $k = n^{\gamma_k}$. We also assume that Υ is γ_s-sparse, that $\mathcal{U}$ stands for a copy of the corresponding clustering statistic under scaling factor Λ, and that Assumption 1 is satisfied.

5.1 Site binning

Ignoring small and large scaling factors We first show that, under Assumption 1, the scaling factor has non-negligible mass between two bounded values.

Proposition 7 (Bounding the scaling factor) *We have*

$$\mathbb{P}\left[\underline{\lambda} \leq \Lambda \leq \overline{\lambda}\right] \geq \chi,$$

where

$$\underline{\lambda} = \frac{g}{M},$$
$$\overline{\lambda} = \frac{2}{1 - e^{-5g}},$$

and

$$\chi = \frac{1 - e^{-5g}}{2}.$$

Proof: From our convention that $\mathbb{E}[\Lambda] = 1$, Markov's inequality implies that

$$\mathbb{P}[\Lambda \geq \overline{\lambda}] \leq \frac{1}{\overline{\lambda}} = \frac{1 - e^{-5g}}{2}.$$

For the other direction, we reproduce the argument in Remark 2. Recall that we assume that

$$\Phi^{-1}\left(e^{-6g}\right) \leq M.$$

Then the probability mass δ below ε (for $\varepsilon < 6g/M$) must be such that

$$\delta \leq e^{-(6g - \varepsilon M)},$$

since

$$e^{-6g} \geq \Phi(M) \geq \delta e^{-\varepsilon M}.$$

Take $\varepsilon = g/M$ so that $\delta \leq e^{-5g}$.

Then we have

$$\mathbb{P}\left[\underline{\lambda} \leq \Lambda \leq \overline{\lambda}\right] \geq 1 - \left(e^{-5g}\right) - \left(\frac{1 - e^{-5g}}{2}\right) = \chi,$$

as desired. ■

Translating the previous proposition into a statement about $\mathcal{U}$-values, we obtain the following.

Proposition 8 (Bounding $\mathcal{U}$-values) *Letting χ be as above, we have*

$$\mathbb{P}\left[\underline{U} \leq U_\Upsilon(\Lambda) \leq \overline{U}\right] \geq \chi,$$

where

$$\underline{U} = e^{-M\overline{\lambda}},$$

and

$$\overline{U} = e^{-2f\underline{\lambda}}.$$

Proof: Recall that

$$U_\Upsilon(\Lambda) = \mathbb{E}[\mathcal{U} \mid \Lambda] = \frac{1}{|\Upsilon|} \sum_{(a,b)\in\Upsilon} e^{-\Lambda d(a,b)}.$$

Since

$$2f \leq d(a,b) \leq M$$

for all $(a,b) \in \Upsilon$, we have

$$e^{-M\Lambda} \leq \frac{1}{|\Upsilon|} \sum_{(a,b)\in\Upsilon} e^{-\Lambda d(a,b)} \leq e^{-2f\Lambda}.$$

The result then follows from Proposition 7. ■

Binning the sites We will now bin the sites whose clustering statistic lie between $\underline{U}$ and $\overline{U}$. The previous proposition guarantees that there is a positive fraction of such sites in expectation. Let

$$\Delta_U = \frac{\gamma_U}{\log n},$$

be the size of the bins in $\mathcal{U}$-space, where $\gamma_U > 0$ is a constant to be fixed later. To avoid taking integer parts, we assume for simplicity that $\overline{U} - \underline{U}$ is a multiple of Δ_U. Let

$$N_U = \frac{\overline{U} - \underline{U} + 2\Delta_U}{\Delta_U},$$

be the number of bins. (The extra $2\Delta_U$ in the numerator accounts for estimation error. See below.) Note that $\overline{U} - \underline{U} = \Theta(1)$ and therefore $N_U = \Theta(\log n)$. We proceed as follows:

- **(Initialization)** For $j = 0, \ldots, N_U$,
  - $\widehat{B}_j = \emptyset$.
- **(Main Loop)** For $i = 1, \ldots, k$,
  - *(Out-of-bounds)* If $\mathcal{U}_i \notin [\underline{U} - \Delta_U, \overline{U} + \Delta_U)$ then $\widehat{B}_0 := \widehat{B}_0 \cup \{i\}$.
  - *(Binning)* Else if

$$\mathcal{U}_i \in [\underline{U} - \Delta_U + (j-1)\Delta_U, \underline{U} - \Delta_U + j\Delta_U)$$

    then $\widehat{B}_j := \widehat{B}_j \cup \{i\}$ and $\widehat{B}_{>0} := \widehat{B}_{>0} \cup \{i\}$.

Restating Proposition 4, we have:

Proposition 9 (Concentration of $\mathcal{U}$-values) *We have*

$$|\mathcal{U}_i - U_\Upsilon(\Lambda_i)| \leq \Delta_U, \qquad \forall i \in \{1, \ldots, k\}, \tag{5}$$

except with probability $\exp\left(-\Omega(n/\log^2 n)\right)$.

Proof: Taking $\zeta = \Delta_U$ in Proposition 4, (5) holds except with probability

$$2n^{\gamma_k} \exp\left(-\Omega(n/\log^2 n)\right) = \exp\left(-\Omega(n/\log^2 n)\right).$$

■

We first show that each bin contains sites with roughly the same scaling factor. We first need a bound on the scaling factors in $\widehat{B}_{>0}$. (Note that, because we needed that the bounds $\underline{U}$ and $\overline{U}$ be independent of Υ (which itself depends on unknown evolutionary distances), Proposition 7 does not apply directly here.)

Proposition 10 (Bounds on selected scaling factors) *Assume (5) holds. There is $\gamma_U > 0$ small enough so that for all $i \in \widehat{B}_{>0}$*

$$\frac{fg}{M^2} \leq \Lambda_i \leq \frac{2M}{f(1 - e^{-5g})}.$$

Proof: Since $U_\Upsilon(\Lambda_i) \leq \overline{U} + 2\Delta_U$ by (5), we have

$$\begin{aligned} \overline{U} + 2\Delta_U &\geq \frac{1}{|\Upsilon|} \sum_{(a,b)\in\Upsilon} e^{-\Lambda_i d(a,b)} \\ &\geq e^{-M\Lambda_i}. \end{aligned}$$

Choose $\gamma_U > 0$ small enough, so that

$$\overline{U} + 2\Delta_U \leq e^{-f\underline{\lambda}}.$$

Taking logarithms,

$$\Lambda_i \geq \frac{fg}{M^2}.$$

Similarly, since $U_\Upsilon(\Lambda_i) \geq \underline{U} - 2\Delta_U$ by (5), we have

$$\begin{aligned} \underline{U} - 2\Delta_U &\leq \frac{1}{|\Upsilon|} \sum_{(a,b)\in\Upsilon} e^{-\Lambda_i d(a,b)} \\ &\leq e^{-2f\Lambda_i}. \end{aligned}$$

Choose $\gamma_U > 0$ small enough, so that

$$\underline{U} - 2\Delta_U \geq e^{-2M\overline{\lambda}}.$$

Taking logarithms,

$$\Lambda_i \leq \frac{2M}{f(1 - e^{-5g})}.$$

■

For $j \in \{1, \ldots, N_U\}$, let

$$U_j = \underline{U} - \Delta_U + (j - 1 + 1/2)\Delta_U,$$

be the midpoint of the j-th bin. Using the fact that $U_\Upsilon(\lambda)$ is strictly decreasing in $\lambda \in \mathbb{R}_+$, we define λ_j as the unique solution to

$$U_\Upsilon(\lambda_j) = U_j,$$

for $j \in \{1, \ldots, N_U\}$.

Proposition 11 (Bin variation) *Assume (5) holds. For any $\gamma_\Lambda > 0$, one can pick $\gamma_U > 0$ small enough (depending on M, f, g) such that for any $j \in \{1, \ldots, N_U\}$ and $i \in \widehat{B}_j$,*

$$|\Lambda_i - \lambda_j| \leq \frac{\gamma_\Lambda}{\log n}.$$

Proof: This follows from Proposition 5. Assume that $U_j \geq U_\Upsilon(\Lambda_i)$. (The other case is similar.) Using the upper bound in Proposition 10 and (5), we get

$$\frac{3}{2}\Delta_U \geq U_j - U_\Upsilon(\Lambda_i) \geq e^{-\Lambda_i M}\left(e^{2f\beta}-1\right) \geq \left(e^{2f\beta}-1\right)\exp\left(-\frac{2M^2}{f(1-e^{-5g})}\right).$$

Hence,

$$\beta \leq \frac{1}{2f}\log\left(1+\frac{3\gamma_U \exp\left(\frac{2M^2}{f(1-e^{-5g})}\right)}{2\log n}\right).$$

■

Next, we argue that at least one bin contains a non-negligible fraction of sites. We say that a bin $\widehat{B}_j$ is *abundant* if

$$\left|\widehat{B}_j\right| \geq k\frac{\chi}{6N_U}.$$

Proposition 12 (Abundant bin) *We have*

$$\exists j^* \in \{1,\ldots,N_U\} \text{ such that } \widehat{B}_{j^*} \text{ is abundant}, \tag{6}$$

except with probability $\exp(-\Omega(n^{\gamma_\delta}))$ *for some* $\gamma_\delta > 0$.

Proof: For the analysis, we introduce *fictitious bins* for the (unknown) expected $\mathcal{U}$-values. That is, for $i = 1,\ldots,k$, we let $i \in B_j$ if

$$U_\Upsilon(\Lambda_i) \in [\underline{U} - \Delta_U + (j-1)\Delta_U, \underline{U} - \Delta_U + j\Delta_U),$$

for some $j \in \{2,\ldots,N_U-1\}$, or $i \in B_0$ otherwise.

Then, there is $j^{**} \in \{2,\ldots,N_U-1\}$ such that

$$\mathbb{P}\left[U_\Upsilon(\Lambda) \in [\underline{U} - \Delta_U + (j^{**}-1)\Delta_U, \underline{U} - \Delta_U + j^{**}\Delta_U)\right] \geq \frac{\chi}{N_U}, \tag{7}$$

that is, the probability that a site falls into bin $B_{j^{**}}$ is at least χ/N_U. This follows immediately from Proposition 8 and the fact that the bins are disjoint and cover the interval $[\underline{U}, \overline{U}]$.

From Lemma 1 and (7),

$$\mathbb{P}\left[|B_{j^{**}}| \leq k\frac{\chi}{2N_U}\right] \leq 2\exp\left(-\frac{(k\chi/2N_U)^2}{2k}\right) = \exp\left(-\Omega(n^{\gamma_k}/\log^2 n)\right).$$

Therefore, if (5) holds, one of $\widehat{B}_{j^{**}-1}$, $\widehat{B}_{j^{**}}$, or $\widehat{B}_{j^{**}+1}$ must contain at least a third of the sites in $B_{j^{**}}$. This occurs with probability at least $1 - \exp(-\Omega(n^{\gamma_\delta}))$ for some $\gamma_\delta > 0$. ■

5.2 Estimating a distorted metric

Estimating evolutionary distances We use an abundant bin to estimate evolutionary distances.

- **(Abundant bin)** Let $\widehat{B}^*$ be any bin with at least $k\frac{\chi}{6N_U}$ sites and set

$$k^* = \left|\widehat{B}^*\right|$$

and

$$\lambda^* = \lambda_j,$$

where j is the index of $\widehat{B}^*$, that is, λ^* is the midpoint of $\widehat{B}^*$.

- **(Evolutionary distances)** For all $a \neq b \in L$, compute

$$\hat{q}^*(a,b) = \frac{1}{k^*} \sum_{i \in \widehat{B}^*} p_\infty^{-1} \left[\mathbb{1}\{\sigma_a^i = \sigma_b^i\} - q_\infty\right].$$

We prove that the $\hat{q}^*(a,b)$ is a good approximation of $e^{-\lambda^* d(a,b)}$.

Proposition 13 (Accuracy) *Let $\gamma_\Lambda > 0, \gamma_q < \gamma_k/2$ be fixed constants. There is a $\gamma_\delta > 0$ such that the following hold except with probability $\exp(-\Omega(n^{\gamma_\delta}))$:*

1. *There is at least one abundant bin. Let $\widehat{B}^*$ be an arbitrary such bin.*
2. *And, for each $i \in \widehat{B}^*$ and for all $a \neq b \in L$,*

$$\left|\hat{q}^*(a,b) - e^{-\lambda^* d(a,b)}\right| \leq \frac{1}{n^{\gamma_q}} + e^{-\lambda^* d(a,b)} \left(e^{\frac{\gamma_\Lambda d(a,b)}{\log n}} - 1\right). \tag{8}$$

Proof: The result follows from Propositions 9, 10, 11 and 12, and the following lemma.

Lemma 3 *Let $\gamma_\Lambda > 0$, $\gamma_q < \gamma_k/2$ be fixed constants. Let*

$$\tilde{k} \geq k\frac{\chi}{6N_U},$$

and

$$\frac{fg}{M^2} \leq \tilde{\lambda}, \tilde{\lambda}_1, \ldots, \tilde{\lambda}_{\tilde{k}} \leq \frac{2M}{f(1-e^{-5g})}$$

be such that, for all i,

$$\left|\tilde{\lambda}_i - \tilde{\lambda}\right| \leq \frac{\gamma_\Lambda}{\log n}. \tag{9}$$

Let $\tilde{\sigma}_L^i \sim \overline{\mathcal{D}}[T, \tilde{\lambda}_i, r]$ independently for all i. Then, for all $a \neq b \in L$,

$$\left|\frac{1}{\tilde{k}}\sum_{i=1}^{\tilde{k}} p_\infty^{-1}\left[\mathbb{1}\{\tilde{\sigma}_a^i = \tilde{\sigma}_b^i\} - q_\infty\right] - e^{-\tilde{\lambda}d(a,b)}\right| \leq \frac{1}{n^{\gamma_q}} + e^{-\tilde{\lambda}d(a,b)}\left(e^{\frac{\gamma_\Lambda d(a,b)}{\log n}} - 1\right).$$

except with probability $\exp\left(-\Omega(n^{\gamma_k - 2\gamma_q}/\log n)\right)$.

Proof: In Lemma 1, take $m = \tilde{k} = \Omega(n^{\gamma_k}/\log n)$, $t = \frac{1}{\tilde{k}}$, and $\zeta = \frac{1}{n^{\gamma_q}}$. Then, except with probability $2n^2 \exp\left(-\Omega(n^{\gamma_k - 2\gamma_q}/\log n)\right)$,

$$\left|\frac{1}{\tilde{k}}\sum_{i=1}^{\tilde{k}} p_\infty^{-1}\left[\mathbb{1}\{\tilde{\sigma}_a^i = \tilde{\sigma}_b^i\} - q_\infty\right] - \frac{1}{\tilde{k}}\sum_{i=1}^{\tilde{k}} e^{-\tilde{\lambda}_i d(a,b)}\right| \leq \frac{1}{n^{\gamma_q}},$$

for all $a \neq b \in L$. Moreover, by (9),

$$\begin{aligned}\left|e^{-\tilde{\lambda}d(a,b)} - \frac{1}{\tilde{k}}\sum_{i=1}^{\tilde{k}} e^{-\tilde{\lambda}_i d(a,b)}\right| &\leq e^{-\tilde{\lambda}d(a,b)}\left|1 - \frac{1}{\tilde{k}}\sum_{i=1}^{\tilde{k}} e^{|\tilde{\lambda}-\tilde{\lambda}_i|d(a,b)}\right| \\ &\leq e^{-\tilde{\lambda}d(a,b)}\left|1 - e^{\frac{\gamma_\Lambda d(a,b)}{\log n}}\right|.\end{aligned}$$

■

■

Tree construction To reconstruct the tree, we use a distance-based method of [DMR09]. We require the following definition.

Definition 7 (Distorted metric [Mos07, KZZ03]) *Let $T = (V, E; L, d)$ be a phylogeny and let $\tau, \Psi > 0$. We say that $\hat{d} : L \times L \to (0, +\infty]$ is a* (τ, Ψ)-distorted metric *for T or a* (τ, Ψ)-distortion *of d if:*

1. [Symmetry] *For all $u, v \in L$, $\hat{d}$ is symmetric, that is,*

$$\hat{d}(u, v) = \hat{d}(v, u);$$

2. [Distortion] $\hat{d}$ *is accurate on "short" distances, that is, for all* $u, v \in L$*, if either* $d(u,v) < \Psi + \tau$ *or* $\hat{d}(u,v) < \Psi + \tau$ *then*

$$\left| d(u,v) - \hat{d}(u,v) \right| < \tau.$$

An immediate consequence of [DMR09, Theorem 1] is the following.

Theorem 3 (See [DMR09].) *Let* $T = (V, E; L, d)$ *be a phylogeny with* n *leaves in* $\mathcal{T}_{f,g}$*. Then topology of* T *can be recovered in polynomial time from a* (τ, Ψ)*-distortion* $\hat{d}$ *of* d *as long as*

$$\tau \leq \frac{f}{5},$$

and

$$\Psi \geq 5g \log n.$$

(The constants above are not optimal but will suffice for our purposes.)

See [DMR09] for the details of the reconstruction algorithm.

We now show how to obtain a $(f/5, 5g \log n)$-distortion with high probability.

Proposition 14 (Distortion estimation) *There are* $\gamma_U, \gamma_\Lambda, \gamma_q, \gamma_k > 0$ *so that, given that the conclusions of Proposition 13 hold, then*

$$\hat{d}(a,b) = -\ln\left(\hat{q}^*(a,b)_+\right), \qquad (a,b) \in L \times L,$$

is a $(\lambda^* f/5, 5\lambda^* g \log n)$*-distortion of* $\lambda^* d$*.*

Proof: Define

$$Ł_2^- = \{(a,b) \in L \times L \,:\, d(a,b) \leq 15g \log n\},$$

and

$$Ł_2^+ = \{(a,b) \in L \times L \,:\, d(a,b) > 12g \log n\},$$

Let $(a,b) \in Ł_2^-$. Note that

$$e^{-\lambda^* d(a,b)} \geq \exp\left(-\left(\frac{2M}{f(1-e^{-5g})}\right) 15g \log n\right) \equiv \frac{1}{n^{\gamma_q'}},$$

where the last equality is a definition. Then, taking γ_q (and hence γ_k) large enough and γ_Λ (and hence γ_U) small enough, from (8) we have

$$\left| \hat{d}(a,b) - \lambda^* d(a,b) \right| \leq \left(\frac{fg}{M^2}\right) \frac{f}{5} \leq \frac{\lambda^* f}{5}.$$

Similarly, let $(a,b) \in \text{Ł}_2^+$. Note that

$$e^{-\lambda^* d(a,b)} < \exp\left(-\left(\frac{fg}{M^2}\right) 12g \log n\right) \equiv \frac{1}{n^{\gamma_q''}},$$

where the last equality is a definition. Then, taking γ_q large enough and γ_Λ small enough, from (8) we have

$$\hat{d}(a,b) \geq 6\lambda^* g \log n \geq 5\lambda^* g \log n + \frac{\lambda^* f}{5}.$$

■

We finally state our main tree-construction result.

Proposition 15 (Tree reconstruction) *Under Assumption 1, given a γ_s-sparse Υ there is a $\gamma_k > 0$ large enough so that the topology of the tree can be reconstructed in polynomial time using $k = n^{\gamma_k}$ samples, except with probability $\exp(-\Omega(n^{\gamma_\delta}))$ for some $\gamma_\delta > 0$.*

Proof: The result follows from Theorem 3 and Proposition 14. ■

Combining Propositions 6 and 15, we get Theorem 2.

6 Concluding remarks

Using techniques from the recent unpublished manuscript [MR11], our results can be extended to handle the more general GTR model of molecular evolution which allows Q-matrices to be time-reversible. This generalization involves choosing pairs of leaves that are not only connected by edge-disjoint paths, but also far enough from each other. One can then use mixing arguments to derive the independence properties required for concentration of the site clustering statistic. We leave out the details.

References

[AAR08] Elizabeth S. Allman, Cecile Ane, and John A. Rhodes. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. *Advances in Applied Probability*, 40(1):228–249, 2008.

[APRS11] Elizabeth S. Allman, Sonja Petrovic, John A. Rhodes, and Seth Sullivant. Identifiability of two-tree mixtures for group-based models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8:710–722, 2011.

[AR06] Elizabeth S. Allman and John A. Rhodes. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *Journal of Computational Biology*, 13(5):1101–1113, 2006. PMID: 16796553.

[Bil95] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1995.

[CH11] Juanjuan Chai and Elizabeth A. Housworth. On Rogers' proof of identifiability for the GTR + Gamma + I model. 2011. Published online at http://sysbio.oxfordjournals.org/content/early/2011/03/27/sysbio.syr023.short.

[Cha96] Joseph T. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.*, 137(1):51–73, 1996.

[CT06] Benny Chor and Tamir Tuller. Finding a maximum likelihood tree is hard. *J. ACM*, 53(5):722–744, 2006.

[DMR09] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Phylogenies without branch bounds: Contracting the short, pruning the deep. In Serafim Batzoglou, editor, *RECOMB*, volume 5541 of *Lecture Notes in Computer Science*, pages 451–465. Springer, 2009.

[EKPS00] W. S. Evans, C. Kenyon, Y. Peres, and L. J. Schulman. Broadcasting on trees and the Ising model. *Ann. Appl. Probab.*, 10(2):410–433, 2000.

[ESSW99] P. L. Erdös, M. A. Steel, L. A. Székely, and T. A. Warnow. A few logs suffice to build (almost) all trees (part 1). *Random Struct. Algor.*, 14(2):153–184, 1999.

[EW04] Steven N. Evans and Tandy Warnow. Unidentifiable divergence times in rates-across-sites models. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 1(3):130–134, 2004.

[Fel04] J. Felsenstein. *Inferring Phylogenies*. Sinauer, Sunderland, MA, 2004.

[HR97] J. P. Huelsenbeck and B. Rannala. Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science*, 276(5310):227–232, 1997.

[KZZ03] Valerie King, Li Zhang, and Yunhong Zhou. On the complexity of distance-based evolutionary tree reconstruction. In *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 444–453, Philadelphia, PA, USA, 2003. Society for Industrial and Applied Mathematics.

[MMS08] Frederick A. Matsen, Elchanan Mossel, and Mike Steel. Mixed-up trees: the structure of phylogenetic mixtures. *Bulletin of Mathematical Biology*, 70(4):1115–1139, 2008.

[Mos07] E. Mossel. Distorted metrics on trees and phylogenetic forests. *IEEE/ACM Trans. Comput. Bio. Bioinform.*, 4(1):108–116, 2007.

[MR95] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Cambridge University Press, Cambridge, 1995.

[MR08] Elchanan Mossel and Sébastien Roch. Detecting and untangling phylogenetic mixtures: An approach based on site clustering. Preprint, 2008.

[MR11] Elchanan Mossel and Sébastien Roch. Phylogenetic mixtures: Concentration of measure in the large-tree limit. Preprint, 2011.

[MS07] Frederick A. Matsen and Mike Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Systematic Biology*, 56(5):767–775, 2007.

[Roc06] Sébastien Roch. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 3(1):92–94, 2006.

[RS10] J. Rhodes and S. Sullivant. Identifiability of large phylogenetic mixture models. Preprint, 2010.

[SS03] C. Semple and M. Steel. *Phylogenetics*, volume 22 of *Mathematics and its Applications series*. Oxford University Press, 2003.

[SS06] M. A. Steel and L. A. Székely. On the variational distance of two trees. *Ann. Appl. Probab.*, 16(3):1563–1575, 2006.

[SSH94] MA Steel, LA Szkely, and MD Hendy. Reconstructing trees when sequence sites evolve at variable rates. *J. Comput. Biol.*, 1(2):153–163, 1994.

[Ste09] Mike Steel. A basic limitation on inferring phylogenies by pairwise sequence comparisons. *Journal of Theoretical Biology*, 256(3):467 – 472, 2009.

[ŠV07a] Daniel Štefankovič and Eric Vigoda. Phylogeny of mixture models: robustness of maximum likelihood and non-identifiable distributions. *J. Comput. Biol.*, 14(2):156–189 (electronic), 2007.

[SV07b] Daniel Stefankovic and Eric Vigoda. Pitfalls of heterogeneous processes for phylogenetic reconstruction. *Syst. Biol.*, 56(1):113–124, 2007.

[WS10] Jihua Wu and Edward Susko. Rate-variation need not defeat phylogenetic inference through pairwise sequence comparisons. *Journal of Theoretical Biology*, 263(4):587 – 589, 2010.